

User Effort and Network Structure Mediate Access to Information in Networks

Jeon-Hyung Kang
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
jeonhyuk@usc.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
lerman@isi.edu

Abstract

Individuals' access to information in a social network depends on its distributed and where in the network individuals position themselves. However, individuals have limited capacity to manage their social connections and process information. In this work, we study how this limited capacity and network structure interact to affect the diversity of information social media users receive. Previous studies of the role of networks in information access were limited in their ability to measure the diversity of information. We address this problem by learning the topics of interest to social media users by observing messages they share online with their followers. We present a probabilistic model that incorporates human cognitive constraints in a generative model of information sharing. We then use the topics learned by the model to measure the diversity of information users receive from their social media contacts. We confirm that users in structurally diverse network positions, which bridge otherwise disconnected regions of the follower graph, are exposed to more diverse information. In addition, we identify user effort as an important variable that mediates access to diverse information in social media. Users who invest more effort into their activity on the site not only place themselves in more structurally diverse positions within the network than the less engaged users, but they also receive more diverse information when located in similar network positions. These findings indicate that the relationship between network structure and access to information in networks is more nuanced than previously thought.

Introduction

People use their social contacts to gain access to information in social networks (Granovetter 1973; Burt 2004), which they can then leverage for personal advantage. However information in social networks is non-uniformly distributed, leading sociologists to explore the relationship between an individual's *network position* and the novelty and diversity of information she receives through her social contacts. Studies of social and organizational networks identified the importance of so-called brokerage positions, which link individuals to otherwise unconnected people (Granovetter 1973; Burt 1995; 2005; Aral and Van Alstyne 2011). By spanning distinct communities, brokerage positions expose

individuals to novel and diverse information, which leads to new job prospects (Granovetter 1973) and higher compensation (Burt 1995; 2004). However, the links that connect individuals in brokerage positions to the rest of the network, generally represent weaker relationships (i.e., acquaintances rather than close friends) (Granovetter 1973; Onnela et al. 2007). The less frequent interactions along these "weak" links limit the amount of information flowing to individuals (Aral and Van Alstyne 2011). Thus, those who are able, and willing, to invest greater effort in social interactions, will manage more connections thereby increasing the volume of information they receive through those links (Aral and David 2012; Miritello et al. 2013b). Specifically, Aral & Van Alstyne (Aral and Van Alstyne 2011) showed that individuals can increase the diversity and novelty of information they receive via email either by placing themselves in brokerage positions, or by communicating more frequently with their social contacts.

In contrast to email and phone interactions, where information is exchanged between a pair of social contacts, social media users broadcast information to all their contacts. Bakshy et al. (Bakshy et al. 2012) showed that weak links collectively deliver more novel information to Facebook users, even though they interact infrequently with these contacts. These findings suggest that an easy way for social media users to increase their access to diverse information is by creating more links, e.g., by following other users. However, cognitive (and temporal) constraints limit an individual's capacity to manage social interactions (Dunbar 1992; Goncalves, Perra, and Vespignani 2011; Miritello et al. 2013b) and process the information they receive (Weng et al. 2012; Hodas and Lerman 2012). In addition, social media users vary greatly in the effort they expend engaging with the site, leading to a large variation in user activity, as measured by the number of messages posted on the site (Wilkinson 2008). The impact of this variation on the information individuals receive and their position in the network is not known. Do users who are able (or at least willing) to be more active on the site receive more diverse information? Do they curate their social links so as to move themselves into network positions that provide more diverse information?

In this work, we use data from the microblogging site Twitter to study the interplay between network structure, the effort Twitter users are willing to invest in engaging

with the site, and the diversity of information they receive from their contacts. Previous studies of the role of networks in individual’s access to information were limited in their ability to measure the diversity of information, using bag-of-words (Aral and Van Alstyne 2011) or predefined categories (Kang and Lerman 2013b) for this task. In this work, we learn topics of interest to social media users from the messages they share with their followers. We present a probabilistic topic model that incorporates human cognitive constraints in a generative model of information sharing and evaluate the model on the task of predicting the messages users retweet. We demonstrate that our model has competitive performance, and unlike other models, it produces descriptions of topics.

We use learned topics to measure the diversity of information users receive from their contacts. This enables us to study the factors that affect the diversity of information in networks. Our findings indicate that the relationship between network structure and access to information is more nuanced than previously thought. First, users cannot increase the diversity of the information they receive by increasing the number of their contacts. Second, we confirm that users in structurally diverse network positions, which bridge otherwise disconnected regions of the follower graph, are exposed to more diverse topics via their contacts than users in less structurally diverse positions. However, we demonstrate that user effort is an important variable mediating access to information in networks. Active users who post more messages on Twitter receive more diverse information even when they are in structurally similar positions to the less active users. This suggests that users who are willing (or able) to engage more on Twitter curate their contacts so as to increase the diversity of the information they receive. Since effort is a useful proxy for individual’s cognitive capacity for (or at least the willingness to invest the time in) processing information in social networks (Miritello et al. 2013a), our work suggests that cognitive factors interact in non-trivial ways with network structure to define access to information in social networks.

Description of Data

Twitter is an online social networking and microblogging service that allows users to follow the activity of others to see the messages they posted or retweeted recently. When a user posts or retweets a message, it is broadcast to all her followers, who are then able to see it in their own streams. Twitter offers an Application Programming Interface (API) for data collection. We used two data sets collected in the past from Twitter. The 2012 data set (Kang and Lerman 2015) contains tweets including a *URL* to monitor information spread over the social network from Nov 2011 to Jul 2012. They start by monitoring potential seed *URLs* containing *http://t.co* from the streaming APIs and collect all tweets containing them. Since the total volume of tweets containing a *URL* is very large, they focus on broadly shared *URLs*. They selected as seeds the *URLs* that appeared more than once in five days from its initial appearance in the streaming APIs based on the heuristic that the *URLs* that have been appeared more often in the streaming APIs will be more pop-

ular on Twitter. They collected the entire history of these seed *URLs* until there were no more tweets containing them within five days from their last appearance in the Twitter REST APIs. This yielded 12.5M tweets with 9.5M users.

The 2014 data set contains the tweets from 5600 initial seed users (Smith et al. 2013) and their friends from Mar 2014 to Oct 2014. Starting with 5,600 initial seed users, they collected all their friends and at least first 200 tweets from their time line. The data set includes 23.8 M tweets from 1.9M users with 17.8M social network links.

Probabilistic Model of User Topics

We use a probabilistic model to learn users’ topics of interest from the messages they share in social media. What information users share, and which messages shared by friends they decide to spread to their followers, depends on a number of factors, such as virality of information being shared, users’ tastes, and their followers’ tastes. To understand information sharing in social networks, social recommendation models (Ma et al. 2008; Wang and Blei 2011; Kang, Lerman, and Getoor 2013) were used to represent users’ interests and items they share by k -dimensional topic vectors. Once these hidden topic vectors are learned from user’s item adoption (i.e., retweeting) history, it is possible to calculate the *personal relevance* of a new item to the user.

We proposed VIP (Kang and Lerman 2015), a model that captures the three basic ingredients of information spread in social media: item’s visibility (v) to a user, its fitness or virality (η), and its (personal) relevance (δ) to the user. While the model improves on previous models, it applies normal distribution assumptions on modeling binary responses, uses full user-item adoption matrix, and provides no descriptions on the learned latent topic space. In this paper, we model binary responses (adopted vs unadopted items) of social media users with multinomial logic model. Stochastic optimization allows us to learn from randomly sampled negative (not adopted) and positive (adopted) dyads without overfitting to the positive ones. Our stochastic inference algorithm handles many user-item dyads and can be distributed for efficient computation. Furthermore, with the help of a probabilistic topic model, we can provide an interpretable low-dimensional representation of information. Figure 1 graphically represents our model.

Item visibility When a user’s message stream is delivered as a list of items, the process of item discovery is biased by the position of each item in the list. A user is more likely to see items near the top of the list than those deeper in the stream (Lerman and Hogg 2014). Hence, items in top stream positions have higher visibility. Since we do not know an item’s exact position, we estimate it as the average visibility of items to user i as follows:

$$v_i \sim \sum_L (\mathbf{G}(1/(1 + \rho_i), L)(1 - \mathbf{IG}(\mu, \lambda, L))) \quad (1)$$

The first factor gives the probability that user i discovers an item depending on the number of items in her stream. The greater the number of new messages user receives between

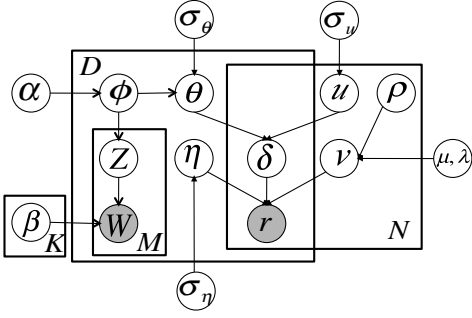


Figure 1: Our model with user topic (u) and item topic (θ) profiles, item’s personal relevance (δ) and visibility to user (v), item fitness (η), expected number of new posts user received (ρ) and item adoption (r). Topic model part has the topic distribution (ϕ) of an item and a distribution (β) over words from a vocabulary of size M . N is the number of users, and D is the number of items.

visits to the site, the less likely the user is to view any specific item. Thus, average visibility depends on the frequency the user visits the site and the rate of posts received. This competition between the rates friends post new messages to the user’s stream and the rate user visits the stream to read the messages modeled by a geometric distribution with success probability $p = 1/(1 + \rho_i)$: $\mathbf{G} = (1 - p)^L p$. The ratio ρ_i of these rates gives the expected number of new messages in a user’s stream. The second factor of gives the probability that user i will navigate to at least $(L + 1)$ -th position in the stream to view the item. This is estimated by the upper cumulative distribution of an inverse gaussian \mathbf{IG} with mean μ and shape parameter λ and variance μ^3/λ :

$$\exp\left(\frac{-\lambda(L - \mu)^2}{2\mu^2 L}\right) \left[\frac{\lambda}{2\pi L^3}\right]^{(1/2)}. \quad (2)$$

Item virality Social media users adopt items even if they had not earlier demonstrated a sustained interest in their topics. This is often the case with viral, general-interest items, such as breaking news or celebrity gossip. Thus, we use “virality” to represent item’s propensity to spread on exposure.

$$\eta_j \sim \mathcal{N}(0, \sigma_\eta^2) \quad (3)$$

Item relevance We calculate personal relevance of an item j to user i as:

$$\delta_{ij} \sim g_\delta(u_i^T \theta_j) \quad (4)$$

where symbol T refers to the transpose operation, u_i represents the topic profile of user i , θ_j represents the topic profile of item j and g_δ is linear function for simplicity.

$$\begin{aligned} u_i &\sim \mathcal{N}(0, \sigma_u^2 I_K) \\ \theta_j &\sim \mathcal{N}(0, \sigma_\theta^2 I_K) \end{aligned} \quad (5)$$

where K is the number of topics.

We use a widely known text mining algorithm Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003),

which analyzes the co-occurrence of the words in documents, to learn the hidden topics representing the documents. In our case, LDA captures the item’s topic distribution ϕ , which is represented as K dimensional vector in the recommendation model. The topic distribution of each document (ϕ_{d_j}) is viewed as a mixture of multiple topics, with each topic (β_k) as a distribution over words. In our setting, the corpus D is a collection of tweet text of the tweet posts. The likelihood of D is computed by multiplying over all documents and all words in each document as follows:

$$p(D|\beta, \phi, z) = \prod_{d_j \in D} \prod_{w \in d_j} \phi_{d_j, z_w} \beta_{z_w, w} \quad (6)$$

where z_w is assigned topic index for each word w in the document d_j , ϕ_{d_j, z_w} is the likelihood of topics z_w for the document d_j and $\beta_{z_w, w}$ is the likelihood of choosing specific word w for the topic z_w .

The generative process for item adoption through a social stream can be formalized as follows:

For each user i
Generate $u_i \sim \mathcal{N}(0, \sigma_u^2 I_K)$
Generate $v_i \sim \sum_L (\mathbf{G}(1/(1 + \rho_i), l)(1 - \mathbf{IG}(\mu, \lambda, l)))$
For each item j
Generate $\eta_j \sim \mathcal{N}(0, \sigma_\eta^2)$
Generate $\phi_j \sim \text{Dirichlet}(\alpha)$
Generate $\epsilon_j \sim \mathcal{N}(0, \sigma_\theta^2 I_K)$ and set $\theta_j = \epsilon_j + \phi_j$
For each word w_{jm}
Generate topic assignment $z_{jm} \sim \text{Mult}(\phi_j)$
Generate word $w_{jm} \sim \text{Mult}(\beta_{z_{jm}})$
For each user i
For each item j on the news feed
Generate the adoption $r_{ij} \sim p(I(r_{ij})|u_i, v, \theta, \eta, O_i)$

Lack of adoption by user i of item j ($r_{ij} = 0$) can be interpreted in two ways: either the user saw the item but did not like it, or the user did not see the item but may have liked it had she seen it. While other models partly account for the lack of knowledge about non-adoptions using smoothing (Wang and Blei 2011; Kang and Lerman 2013a), we properly model visibility of items to users.

We model the user-item adoption with Softmax function, which makes the values of the K dimensional vectors in $[0, 1]$ range. The equation is as follows:

$$p(I(r_{ij})|u_i, v, \theta, \eta, O_i) = \frac{\exp(v_i g_r(\delta_{ij} + \eta_j))}{\sum_{l \in O_i} \exp(v_i g_r(\delta_{il} + \eta_l))} \quad (7)$$

where $I(r_{ij})$ is the indicator function, $I(r_{ij}) = 1$ when user i adopted item j and 0 otherwise, and O_i is the observed items by user i . We define g_r as linear functions for simplicity.

The main objective function is:

$$\begin{aligned}
\ell = & -\frac{1}{2\sigma_u^2} \sum_i^N u_i^T u_i - \frac{1}{2\sigma_\eta^2} \sum_j^D \eta_j^T \eta_j \\
& - \frac{1}{2\sigma_\theta^2} \sum_j^D (\theta_j - \phi_j)^T (\theta_j - \phi_j) \\
& + \sum_i^N \log \left(\sum_l^L (1/\rho_i + 1)(\rho_i/\rho_i + 1)^l (1 - \mathbf{IG}(\mu, \lambda, l)) \right) \\
& - \sum_i^N \sum_j^D \left(\log \left(\sum_{l \in O_i} \exp(v_i(\delta_{il} + \eta_l)) \right) - v_i(\delta_{ij} + \eta_j) \right)
\end{aligned} \tag{8}$$

The last term of the equation minimizes the error between the binary rating and the predicted rating. The second line of the equation minimizes the error between the topics that explain the recommendation and the content. The importance between these two components can be controlled with σ_θ . MAP estimation is equivalent to maximizing the complete log likelihood (ℓ) of U, V, θ, η, ϕ and r given $\sigma_u, \sigma_\theta, \sigma_\eta, \mu, \lambda$ and ρ .

Model Learning

To optimize Eq. (8), we develop a stochastic gradient descent algorithm. Given a current estimate, we take the gradient of Eq. (8) with respect to u_i, θ_j , and η_j and iteratively optimize the parameters $\{u_i, \theta_j, \eta_j\}$. Derived update equations are:

Algorithm 1 Stochastic Optimization

```

Initialize model parameter  $U, V, \theta, \eta, \phi, \nabla$ 
for  $t = 1$  to  $T$  do
  for  $u$  in  $U$  do
    Choose random  $|r_i|$  mini batch  $S_i$  from  $D-r_i$ 
    Generate  $O_i = r_i \cup S_i$ 
    for  $j$  in  $O_i$  do
       $u_i \leftarrow u_i - \mu [v_j \theta_j \nabla + \frac{1}{2|r_i| \sigma_u^2} u_i]$ 
       $\theta_j \leftarrow \theta_j - \mu [v_i u_i \nabla + \frac{1}{2|r_j| \sigma_\theta^2} (\theta_j - \phi_j)]$ 
       $\eta_j \leftarrow \eta_j - \mu [v_i \nabla + \frac{1}{2|r_j| \sigma_\eta^2} \eta_j]$ 
    end for
  end for
end for

```

where $|r_i|$ is the number of items adopted by user i and $|r_j|$ is the number of users who adopted item j . We generate a set of observed items O_i by adding randomly sampled $|r_i|$ number of items from the unadopted set ($D-r_i$) and incrementally learning from the unadopted and adopted item set of each user. We use the learning rate μ with discount by a factor of 0.9 in each iteration (Koren, Bell, and Volinsky 2009).

The equation for gradient (∇) is as follows:

$$\nabla = \frac{\exp(v_i g_r(\delta_{ij} + \eta_j))}{\sum_{l \in O_i} \exp(v_i g_r(\delta_{il} + \eta_l))} - I(r_{ij}). \tag{9}$$

Table 1: Model parameters used in this study.

Parameters	Value
number of topics	$K=100$
user topic profile	$\sigma_u^2=10^4$
item topic profile	$\sigma_\theta^2=10^4$
item fitness	$\sigma_\eta^2=10$
law of surfing	$\mu = 14.0$ $\lambda = 14.0$
views per post	38
typical posting rates	1.4

The proposed recommendation model can be updated incrementally to model dynamic user adoptions in real time. It is also computationally efficient since it can be distributed by decomposing the data set over multiple computers.

Model Selection

We use the same ‘‘law of surfing’’ parameters, $\mu = 14.0$ and $\lambda = 14.0$, as (Kang and Lerman 2015; Hogg, Lerman, and Smith 2013; Hogg and Lerman 2012) did in their study of social media. The expected number of new posts including a URL user i received, ρ_i , is computed by $rate_i^{(url \text{ posts received})} / rate_i^{(visits)}$. The rate $rate_i^{(posts received)}$ is proportional to the number of friends ($N_{frd(i)}$) i follows and their average posting frequency. To estimate posting frequency of all users, we use the typical URL posting rates of users from our data: $rate_i^{(posts received)} = 1.4 * N_{frd(i)}$. We estimate user i ’s visiting rate ($rate_i^{(visits)}$) using the number of posts of user i ($N_{posts(i)}$). (Hogg, Lerman, and Smith 2013) estimated that average number of visits per post was 38 (2014 data set) for Twitter users. Also, since around 20% of tweets include a URL (Chaudhry et al. 2012), the posting rate of user i becomes $rate_i^{(visits)} = 7.6 * N_{posts(i)}$ (2012 data set).

For the model hyper-parameters, we vary the parameters $K \in \{10, 30, 50, 100, 200\}$, and $\{\lambda_u, \lambda_\theta\} \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$ by using grid search on validation set. Throughout this paper, we set parameters $K = 100, \lambda_u = 0.01, \lambda_\theta = 0.001$, both for PMF and CTF that performed the best for PMF. For the fitness parameter of VIP (Kang and Lerman 2015) and the proposed model, we vary $\sigma_\eta^2 \in \{10^{-4}, 10^{-3}, \dots, 10^4\}$, while we fix other parameters: $\sigma_\theta^2 = 10^4$ and $\sigma_u^2 = 10^4$. In this paper, we set $\sigma_\eta^2 = 10$.

Model Evaluation

We evaluate the proposed model by using it to predict which items users will adopt. For this task, user i ’s adoption of item j shared by a friend is obtained by point estimation with optimal variables $\{\theta^*, u^*, v^*, \eta^*\}$:

$$\begin{aligned}
\mathbb{E}[r_{ij} | \mathcal{D}] & \approx \mathbb{E}[v_i | \mathcal{D}]^T (\mathbb{E}[\delta_{ij} | \mathcal{D}] + \mathbb{E}[\eta_j | \mathcal{D}]) \\
r_{ij}^* & \approx v_i^* (u_i^{*T} \theta_j^* + \eta_j^*)
\end{aligned} \tag{10}$$

where \mathcal{D} is the training data. The adoption probability is decided by user visibility v_i^* , user topic profile u_i^* , item topic profile θ_j^* , and item fitness η_j^* .

To evaluate the performance, we use precision (P), recall (R) and normalized discounted cumulative gain (nDCG) for top-x recommended posts.

Table 2: Overall prediction performance comparison using Precision@ x ($P@x$), Recall@ x ($R@x$), normalized DCG@ x ($nDCG@x$) on Twitter dataset.

Model	Text	P@10	R@10	nDCG@10
Random	No	0.0483	0.3738	0.2410
Fitness	No	0.0798	0.5924	0.3630
Relevance	No	0.0647	0.4383	0.3170
VIP	No	0.0984	0.6446	0.4205
Softmax-CTR	Yes	0.1047	0.6105	0.4123
Our Model	Yes	0.1138	0.7022	0.4619

$P@x$ computes the fraction of items that are adopted by each user in top- x items in the list. We average the precision@ x of all users.

$R@x$ computes the fraction of adopted items that are successfully discovered in top- x ranked list out of all adopted items by each user. We average the recall@ x of all users.

$nDCG@x$ computes the weighted score of adopted items based on the position in the top- x list. It penalizes adopted items in the bottom of the top- x list. We average the nDCG@ x of all users.

We divide each user’s adopted items into five folds and construct the training set and the test set. We use five-fold cross validation and compare performance of the proposed model to five baseline models: RANDOM, FITNESS, RELEVANCE, VIP, CTR. The RANDOM baseline chooses items at random from among the items in user i ’s stream, i.e., items adopted by i ’s friends. The baseline FITNESS uses item fitness values (η) learned by VIP to recommend k highest fitness items. The baseline RELEVANCE bases its recommendations on user-topic and item-topic vectors learned by PMF. Collaborative Topic Regression (CTR) (Wang and Blei 2011) was originally introduced to recommend scientific articles. It combines collaborative filtering (PMF) and probabilistic topic modeling (LDA). It captures two K -dimensional lower-rank user and item hidden variables from user-item adoption matrix and the content of the items. This model uses textual information and negative dyads, but unlike our method it uses ℓ_2 function instead of a Softmax. Here for a fair comparison, we implemented a Softmax version. Based on our experiment Softmax-CTR outperformed original CTR due to the binary adoptions of social media.

Table 2 shows the models’ overall performance on the user–item adoption prediction task. In this paper, we set $x=10$ since recommending too many items is not realistic. From our experiments, we found that results are consistent with different number of k . While nDCG@ x uses the position of correct answer in the top- x ranked list, it does not penalize for unadopted items or missing adopted items in the top- x ranked list, therefore one has to consider the performance of all three metrics together. Intuitively a better model should have higher $P@x$, $R@x$, and nDCG@ x .

The experimental results show that the proposed model dramatically outperforms the random model with 135.61% and 87.85% respectively on precision and on recall. A comparison against the random model is important to uncover the complexity of the post-recommendation task. FITNESS and RELEVANCE models yield 62.21% and 33.95% improvement over the random model in terms of precision,

Var.	Description
S_i	number of active friends
N_{D_i}	network diversity
O_i	avg. vol. of outgoing info. (# tweets/day)
u_i	user-topic vector. (k -dimensional vector)
F_{TD_i}	friend topic diversity

Table 3: Variables used in the study.

and 58.48% and 17.25% in terms of recall respectively. The gain of VIP over RELEVANCE is 52.08% on precision and 47.06% on recall, while the one of CTR over RELEVANCE is 61.82% on precision and 39.28% on recall. This shows that accounting for cognitive biases dramatically improves predictability of user item adoptions in social media as much as accounting for text description of items alone. Among all models, the proposed model yields best performance, showing that modeling text, as well as visibility, is critical in social media recommendation.

Information Access in Networks

We use the topics learned by the proposed model to study how information is distributed in a network and what users can do to increase the diversity of information they receive from their social media friends. In order to use the messages users posted, in addition to friends’ messages they retweeted, we changed the model by assigning visibility equal to one to each original message user posted.

Definition of Variables

Following (Aral and Van Alstyne 2011; Aral and David 2012) we define a set of variables we use to characterize users, their network position, and information diversity.

Network size We define the network size S_i of user i as the number of friends from whom user i received messages during a time period Δt , which we take to be the data collection period. We only consider active friends, i.e., friends who posted messages during Δt . Network size is defined as

$$S_i = \sum_{l \in N_i^{frd}} I(r_l) \quad (11)$$

where N_i^{frd} is the set of friends of user i and the indicator function $I(r_l)$ is one if and only if friend l tweeted during the time period Δt and zero otherwise.

Network diversity User’s position in a network significantly impacts the diversity of received information. Position can be characterized by its structural diversity, which represents how many otherwise unconnected contacts user i has. We measure structural diversity of a network position using local clustering coefficient (Watts and Strogatz 1998), C_i , which quantifies how often user i ’s contacts are linked (regardless of the direction of the link):

$$C_i = \frac{2 \times |\{e_{jk} : j, k \in N_i^{frd}, e_{jk} \in E\}|}{S_i(S_i - 1)} \quad (12)$$

Table 4: Keywords associated with the top 10 topics of users in different positions within the network. Users are divided into two populations based on their network diversity (ND).

#	Users in a Low ND	Users in a High ND
1	lesson weight loss acoustic lose motive guitar flash gain	profession connect profile webdesign bigdata update
2	pet dog animal adopt praise cat rescue love mate relax	children parent surgery inch anxiety obesity autism
3	read book review kindle novel cover publish buddha	united kingdom stadium arena holland yankees
4	good happy hope morn birthday wish love like	prosecute labour governor palestinian nationwide peru
5	yoga workout exercise jump doctor fit body back diet	ferguson pray brooklyn documentary oakland
6	graphic japanese poetry manga cinema photo	art center science exhibit culture paper draw museum
7	oil kale gene napa sausage wrap aspire coal trainer	camera shoot timeline canon len accent timeline possess
8	children parent common journey ready pack escape	worldcup shout football soccer illinois player sold
9	home design studio site interior built lawn layout	space mars nasa planner newton isaac modern
10	beauty summer city park resort nation beach island	free win get email gift chance enter offer ticket

The variable $e_{jk} = 1$ if user j follows user k or vice versa; otherwise, $e_{jk} = 0$. The total number of possible connections among contacts is $S_i(S_i - 1)$. High clustering coefficient implies low network diversity, and vice versa. Therefore, we define network diversity of user i as $ND_i = 1 - C_i$. Note that brokerage positions have high network diversity, while individuals in tightly-knit communities have are in positions with low network diversity.

User effort Most social media sites, including Twitter, display items from friends as a chronologically ordered list, with the newest items at the top. A user scans the list and if she finds an item interesting, she may share it with her followers by retweeting it. She will continue scanning the list until she loses interest or distracted (Hodas and Lerman 2012). It is difficult to quantify how much of the list a user processes, since the site does not provide this information. Instead, we use user *activity* as a heuristic for the effort users are willing (or able) to invest in Twitter. We measure user i 's activity by the average number of messages the user tweets and retweets per day:

$$O_i = \frac{|r_i|}{\Delta t} \quad (13)$$

where $|r_i|$ is the number of tweets from user i .

Friend topic diversity We measure the diversity of information user i receives from friends by the the variance of friends' topic interests: when most of friends have distinct, non-overlapping, interests, topic diversity will be high, whereas when most of friends have similar topic interests it will be low. We define friend topic diversity as the average pair-wise cosine distance of friends' topic interest vectors.

$$FTD_i = \frac{2 \times \sum_{j \in N_i^{f_{rd}}} \sum_{k \in N_i^{f_{rd}}} (1 - \text{Cos}(u_j, u_k))}{S_i(S_i - 1)} \quad (14)$$

Information and Network Structure

Information is not uniformly distributed in a network: users in brokerage positions are interested in systematically different topics than users within denser communities. To study user-topic distribution, we rank users according to network diversity (ND) and split them into two equal sized groups: high and low network diversity. Table 4 compares the representative keywords of the top ten topics from the topic profiles of users in these two groups. Users in high network diversity positions tend to be interested in more general topics, such as sports (“worldcup”, “yankees”, “lad”), current events (“ferguson”, “oakland”), business (“profession”, “big data”), health (“surgery”, “obesity”), politics (“peru”, “palestinian”), arts (“art”, “exhibit”, “camera”), science (“science”, “nasa”, “space”), promotion (“gift”, “offer”), etc. According to sociological theory, users in such brokerage positions spanning multiple unconnected communities are exposed to diverse information (Burt 1995); therefore, it makes sense that the topics they have in common are the more general topics. On the other hand, users in positions of low network diversity focused on more specialized topics, such as hobbies (“guitar”, “book”, “yoga”, “manga”), pets (“dog”, “cat”), family (“birthday”, “children”), food (“oil”, “kale”), vacation (“journey”, “escape”, “island”), home & garden (“home”, “interior”).

Increasing Exposure to Diverse Information

How can users increase the amount of diverse information they receive in social media? Do they follow more people to increase the volume of information received? Or do they move themselves into special network positions? To examine how user effort affects information access, we split users into four classes based on the average number of tweets they post daily (O). The top quartile contains the most active users, who post more than 5.3 tweets per day, the second quartile contains users who post from 3.1 to 5.3 tweets per day and the third and the bottom quartile contains from 1.9 to 3.1 and fewer than 1.9 tweets per day respectively.

Figure 2 shows the relationship between diversity of received information, measured by friend topic diversity (FTD), and user's network size (S), for these classes of Twitter users. The trends among these four classes of users are somewhat different, indicating that people use different strategies to access information in network. Active users who expend more effort on Twitter (red circles in Figure 2) increase their exposure to diverse information by adding more friends (0.1874, $p < .01$). However, when the bottom quartile users (blue squares in Figure 2) add friends, this actually decreases the diversity of information they are exposed to until around 100 friends. After that point, information diversity slowly increases. For the same network size, the less active users actually receive more diverse information than the more active user until around 100 friends. Apparently, network size itself cannot provide an access to diverse information (when $S > 100$) since the network structure can vary significantly.

In addition to network size, network position is known to play an important role in determining access to information.

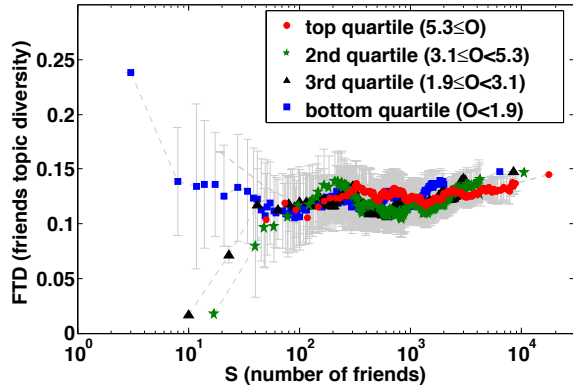


Figure 2: Diversity of received information as a function of user’s network size. Users are divided into four populations based on their effort: red circles represent the more active users, (who post more than 5.3 tweets per day on average), green stars represent the 2nd quartile ($3.1 \leq O_i < 5.3$), black triangles represent 3rd quartile ($1.9 \leq O_i < 3.1$) and the blue squares represent that bottom quartile users (who post fewer than 1.9 tweets per day on average). We discretize values into equal-sized bins for each quartile.

In social and email communication networks, people in high network diversity positions receive more novel and diverse information (Granovetter 1973; Aral and Van Alstyne 2011; Aral and David 2012). We tested whether the same conclusions hold for Twitter using topics learned by the proposed model. Figure 3 shows the relationship between friend topic diversity (FTD_i) and structural network diversity (ND_i) for the four classes of users divided according to their effort. There is a strong correlation (0.9212 ($p < .01$)) for bottom quartile users (blue squares in Figure 3), between network position and information diversity, correlation values decrease with increasing user effort (3rd quartile 0.9162 ($p < .01$) and 2nd quartile 0.7774 ($p < .01$)). When these users place themselves in more structurally diverse position within the Twitter network, they receive on average more topically diverse tweets from friends than users who place themselves in less structurally diverse network positions. However, the correlation between FTD and ND for active users (red circles in Figure 3) is far less, 0.3248 ($p < .01$). These users are generally exposed to more diverse information than the less active users, regardless of their network position. Also, active users in low network diversity positions receive more diverse information than the less active users in similar positions. These results demonstrate that the effort users are willing to invest in using social media is an important factor in access to diverse information.

Why are highly active users exposed to more diverse information? To address this question, we study how network diversity changes as users add more friends. Figure 4 shows this relationship for users separated into two classes based on their activity or effort. Overall, network diversity increases with network size (after around 100 friends), which is not surprising since probabilistically as the number of people in a network grows, any two people are less likely to be

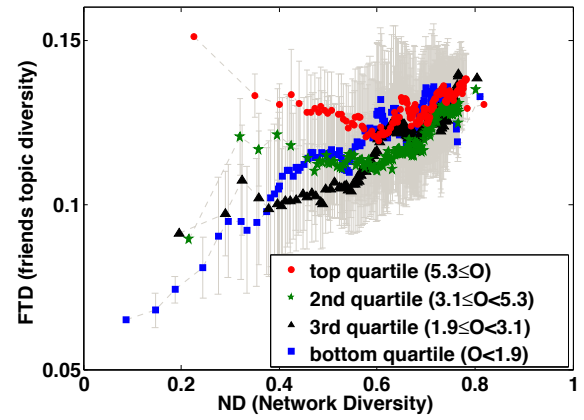


Figure 3: Friend topic diversity (FTD_i) of a user as a function of the network diversity (ND_i) in the 2014 Twitter data set. We show the average of FTD_i for the same network diversity (ND_i) users with their standard deviation ranges in grey color. Users in the higher network diversity positions tend to be exposed to more diverse information, with active users receiving more diverse information regardless of their position in the network structure. We group ND values into equal-sized bins and compute the mean of both ND and FTD within each bin.

connected to each other. Active users overall place themselves in more structurally diverse positions.

Surprisingly, network diversity initially decreases with network size for both user populations, reaching a minimum around $S = 100$. A potential explanation of this effect involves the Dunbar number. Dunbar (Dunbar 1992) argued that finite human cognitive capacity constrains the number of social interactions individuals can manage, limiting size of social groups to about 100–200 individual. Research has validated the impact of cognitive constraints on online social interactions (Goncalves, Perra, and Vespignani 2011; Kang and Lerman 2013b). Similar arguments could apply to our setting. Minimum network diversity corresponds to maximal social connectivity, which in our Twitter data set occurs when users have around 100 friends. While their social networks can grow beyond that size, increasing network diversity implies that new friends are less likely to form a community.

The minimum in network diversity for the less active users occurs at lower values than for the more active users. This suggests that active users who invest more effort into using Twitter can manage larger communities of connected friends than the less-active users. This observation is in line with cognitive limits on social interactions theory: users who have a greater capacity for social interactions (or who may simply be willing to invest more time and effort in social interactions) will have more interactions on Twitter (higher activity), and they will also tend to belong to larger social groups (higher network size), simply because they are better capable of managing their social connections. At this time we cannot prove this intriguing possibility, and leave it as a question for future research.

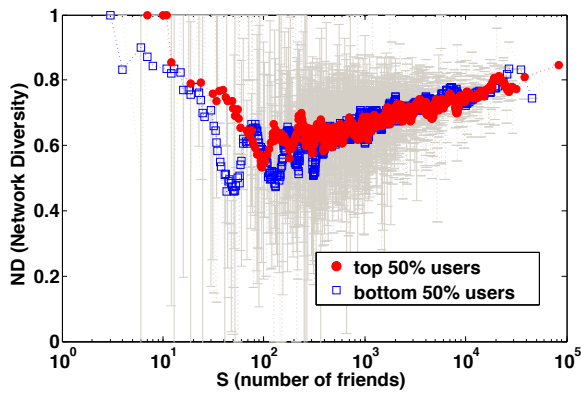


Figure 4: Network diversity (ND) as a function of the number of active friends (S) in the 2014 Twitter data set. We use equal-sized bins for each class.

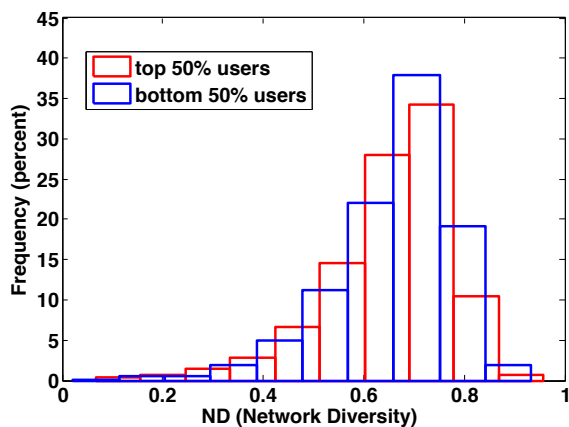


Figure 5: Histograms of network diversity (ND) of users in the 2014 Twitter data set. Users are divided into two populations based on their effort (O). The peak of top 50% users is higher than bottom 50% users, while bottom 50% users tend to have higher ND .

Related Work

A pair of classic theories has linked an individual’s position within a network to the novelty and diversity of information she receives through her social contacts. The theoretical argument, known as “the strength of weak ties” (Granovetter 1973), explored the relationship between social links and the information people receive along those links. Specifically, the weak links, representing infrequent social interactions, were shown to deliver novel information to people, providing new social and economic opportunities (Uzzi 1997; Reagans and Zuckerman 2001; Reagans and McEvily 2003; Allen 2003).

Burt (Burt 1995; 2004; 2005) argued that weak ties act as bridges between different communities. Individuals with many such ties are in what he termed “brokerage positions” in the network, which allows them with access, and benefit from, novel information residing in diverse sources. Empirical research on mobile phone (Onnela et al. 2007), email

communication (Aral and Van Alstyne 2011; Iribarren and Moro 2011), and online social networks (Grabowicz et al. 2011; Centola and Macy 2007; Centola 2010) supported the weak ties arguments about the nature of interactions on a network and its structure.

Aral & Van Alstyne show that both structurally diverse brokerage positions in the network and high frequency communication along social ties provided access to diverse and novel information in the email communication network. In social media, Kang & Lerman (Kang and Lerman 2013b) showed that increasing activity of social media friends a user follows affected how much novel information user received from them, while increasing network diversity provided access to more topically diverse information, but not the other around. Bakshy et al. (Bakshy et al. 2012) showed that, although strong ties are individually more influential, weak ties increased the diversity of information received.

Cognitive constraints on social interactions provide an interesting perspective on the structure and function of social networks. Dunbar argued that people have a limited ability, defined by their brain’s capacity, to manage social interactions, which gives rise to maximum social group size (Dunbar 2003). Although social media was believed to expand the size of human social networks, research showed that the maximum number of friends that Twitter users interact with is around 100-200 (Goncalves, Perra, and Vespignani 2011), similar to the Dunbar number. Cognitive constraints could also explain the findings of (Aral and Van Alstyne 2011; Aral and David 2012), namely that cognitive constraints create a trade-off between the complexity of social interactions (given by network diversity) and the intensity of interactions along structurally complex links, resulting in “diversity-bandwidth trade-off.” Unlike previous researchers, we examined how users vary in their capacity for social interactions (or activity), and how this capacity defines their level of engagement with the social media site and access to diverse information.

Recommender system (Herlocker et al. 1999; Sarwar et al. 2001; Karypis 2000) examines item ratings of many people to discover their preferences and recommend new items that were liked by similar people. Latent-factor models, such as probabilistic matrix factorization (Salakhutdinov and Mnih 2008; Koren, Bell, and Volinsky 2009; Wang and Blei 2011), have shown promising in creating better recommendations by incorporating personal relevance into the model. Many social recommender systems have been proposed by matrix factorization techniques for both user’s social network and their item rating histories (Ma et al. 2008). In addition to modeling user-item adoptions, researchers integrate social correlation between users (Purushotham, Liu, and Kuo 2012), topic influences of friends (Kang and Lerman 2013a), and cognitive biases (Kang and Lerman 2015) in social recommender system.

Recommender systems often focus on understanding user preferences based on the history of observed actions to recommend possible future likes and interests. One of the key challenge is how to increase the variety of recommended items without the expenses of sacrificing the accuracy. The trade-off between exploration and exploitation is important

to prevent over-specialization where we never recommend items outside of the history of user's actions. Most of the current approaches focus on proposing new intra-list diversity metrics (Ziegler et al. 2005; Agrawal et al. 2009) to diversify recommendations. Our study shows that users increase activity to access diverse information. We can estimate how much user opens to diverse information by taking into account the engagement levels as well as the network diversity of the user.

Conclusion

The idea that network structure affects the novelty and diversity of information people receive from their social contacts has long fascinated sociologists (Granovetter 1973; Burt 1995). However, humans also have a finite cognitive capacity, which constraints how many social relations they are able to manage (Dunbar 1992). The interplay between network structure and cognitive constraints has important implications for how people gain access to information in social networks in general, and on social media in particular. In this paper, we explored these questions using data from a popular social media platform Twitter, where users create links in order to receive information, in the form of short text messages called tweets, from other people.

One of the challenges we faced is measuring the diversity of information users receive from their friends on Twitter. We addressed this challenge by using a probabilistic model to learn users' topics of interest from the messages they receive and share on Twitter. Our model incorporates the text of messages and a user's network in a generative model of information spread. We then used learned topics to measure diversity of the information a user is exposed to as the variance of topic interests of the user's friends.

By quantifying information diversity, we can study the factors that affect information access in networks. We confirmed that network position plays an important role: users can increase the amount of diverse information they receive by increasing the structural diversity of their network position, rather than simply increasing the number of people they follow. However, we also identified user effort as an important factor mediating access to information in networks. Users who post (and consume) more messages place themselves in positions of higher network diversity than the less active users. Even when they are in structurally similar positions, the more active users receive more diverse information. This suggests that users who invest greater effort into using Twitter may have higher cognitive capacity for processing information, or they may simply be able to devote more time to such interactions (Miritello et al. 2013b). These users curate their links so as to increase the diversity of information they receive. One mechanism for accomplishing this is to break links so as to reduce the redundancy of received information. Even when these actions do not change a user's structural position within the network, they serve to increase information diversity. Our work underscores the importance of cognitive factors and variation in effort in access to information in networks. Work is needed to further disentangle these factors.

Acknowledgments

This work was supported in part by AFOSR (contract FA9550-10-1-0569), by DARPA (contract W911NF-12-1-0034), and by the NSF (under grants CIF-1217605 and SMA-1360058).

References

- Agrawal, R.; Gollapudi, S.; Halverson, A.; and Jeong, S. 2009. Diversifying search results. In *WSDM*. ACM.
- Allen, T. 2003. Managing the flow of technology: Technology transfer and the dissemination of technological information within the r&d organization. *MIT Press Books* 1.
- Aral, S., and David, V. 2012. The anatomy & dynamics of vision advantages. In *ICIS*.
- Aral, S., and Van Alstyne, M. W. 2011. The Diversity-Bandwidth tradeoff. *American Journal of Sociology* 117(1):90–171.
- Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *WWW*.
- Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.
- Burt, R. 1995. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press.
- Burt, R. S. 2004. Structural holes and good ideas. *The American Journal of Sociology* 110(2):349–399.
- Burt, R. S. 2005. *Brokerage and closure: An introduction to social capital*. Oxford University Press.
- Centola, D., and Macy, M. 2007. Complex contagions and the weakness of long ties1. *American Journal of Sociology* 113(3):702–734.
- Centola, D. 2010. The spread of behavior in an online social network experiment. *science* 329(5996):1194–1197.
- Chaudhry, A.; Glodé, L. M.; Gillman, M.; and Miller, R. S. 2012. Trends in twitter use by physicians at the american society of clinical oncology annual meeting, 2010 and 2011. *Journal of Oncology Practice* 8(3):173–178.
- Dunbar, R. I. M. 1992. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution* 22(6):469–493.
- Dunbar, R. 2003. Evolution of the social brain. *Science* 302(5648):1160–1161.
- Goncalves, B.; Perra, N.; and Vespignani, A. 2011. Modeling users' activity on twitter networks: validation of dunbar's number. *PLoS One* 6(8):e22656.
- Grabowicz, P. A.; Ramasco, J. J.; Moro, E.; Pujol, J. M.; and Eguluz, V. M. 2011. Social features of online networks: the strength of weak ties in online social media. *CoRR* abs/1107.4009.
- Granovetter, M. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78(6):1360–1380.
- Herlocker, J.; Konstan, J.; Borchers, A.; and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering. In *SIGIR*.

- Hodas, N. O., and Lerman, K. 2012. How limited visibility and divided attention constrain social contagion. In *SocialCom*.
- Hogg, T., and Lerman, K. 2012. Social dynamics of digg. *EPJ Data Science* 1(5).
- Hogg, T.; Lerman, K.; and Smith, L. M. 2013. Stochastic models predict user behavior in social media. In *SocialCom*.
- Iribarren, J., and Moro, E. 2011. Affinity paths and information diffusion in social networks. *Social networks* 33(2):134–142.
- Kang, J.-H., and Lerman, K. 2013a. LA-CTR: A limited attention collaborative topic regression for social media. In *AAAI*.
- Kang, J.-H., and Lerman, K. 2013b. Structural and cognitive bottlenecks to information access in social networks. In *Hypertext*.
- Kang, J.-H., and Lerman, K. 2015. VIP: Incorporating human cognitive biases in a probabilistic model of retweeting. In *SBP*.
- Kang, J.-H.; Lerman, K.; and Getoor, L. 2013. LA-LDA: A limited attention model for social recommendation. In *SBP*.
- Karypis, G. 2000. Evaluation of item-based top-n recommendation algorithms. Technical report, DTIC Document.
- Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.
- Lerman, K., and Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PLoS One* 9(6):e98914.
- Ma, H.; Yang, H.; Lyu, M.; and King, I. 2008. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*, 931–940. ACM.
- Miritello, G.; Lara, R.; Cebrian, M.; and Moro, E. 2013a. Limited communication capacity unveils strategies for human interaction. *Scientific Reports* 3.
- Miritello, G.; Moro, E.; Lara, R.; Martínez-López, R.; Belchamber, J.; Roberts, S. G. B.; and Dunbar, R. I. M. 2013b. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks* 35(1):89–95.
- Onnela, J.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaski, K.; Kertész, J.; and Barabási, A. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104(18):7332–7336.
- Purushotham, S.; Liu, Y.; and Kuo, C. 2012. Collaborative topic regression with social matrix factorization for recommendation systems. In *ICML*.
- Reagans, R., and McEvily, B. 2003. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative Science Quarterly* 48(2):240–267.
- Reagans, R., and Zuckerman, E. 2001. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science* 12(4):502–517.
- Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. *NIPS* 20:1257–1264.
- Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Itembased collaborative filtering recommendation algorithms. In *WWW*.
- Smith, L. M.; Zhu, L.; Lerman, K.; and Kozareva, Z. 2013. The role of social media in the discussion of controversial topics. In *SocialCom*.
- Uzzi, B. 1997. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly* 35–67.
- Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*.
- Watts, D., and Strogatz, S. 1998. The small world problem. *Collective Dynamics of Small-World Networks* 393:440–442.
- Weng, L.; Flammini, A.; Vespignani, A.; and Menczer, F. 2012. Competition among memes in a world with limited attention. *Scientific Reports* 2.
- Wilkinson, D. M. 2008. Strong regularities in online peer production. In *EC*, 302–309.
- Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving recommendation lists through topic diversification. In *WWW*, 22–32. ACM.